Elaborato 2: Programmazione di sistema (I/O, fork, exec, signal, pipe, fifo code di messaggi, memoria condivisa, semafori).
Consegna: entro il 15 Giugno 2010 ore 9:00.
Modalità di consegna:

1) Rinominare il file contenente l'elaborato con il proprio numero di matricola. Si ricorda che la consegna è individuale, pertanto ogni studente dovrà consegnare una copia dell'elaborato.
2) Riportare in calce al file contenente l'elaborato un commento che includa: matricola, nome e cognome, data di consegna, titolo dell'elaborato.
3) Fare l'upload del file su http://amarena.sci.univr.it/
    a. Seguire i link: Accesso pubblico → Laboratorio Sistemi Operativi 2010 → "Nome_docente_del_corso"
    b. A questo punto dovreste trovarvi all'interno di anonymous / Laboratorio Sistemi Operativi 2010 / Nome_docente_del_corso
    c. Cliccare sulla freccia alla destra della voce Elaborato 2 (sotto la colonna Azione), quindi su Nuovo → Documento
    d. Compilare i campi del form che appare inserendo il file di cui si vuole fare l'upload in "File locale", il vostro nome, cognome e n° di matricola su "Nome del documento".
    e. Premere OK
4) Si ricorda inoltre che non si potranno né modificare né visualizzare i file di cui è stato fatto l'upload.
5) Per qualunque problema durante la sottomissione dell'elaborato contattare il docente del relativo corso (Bombieri per Informatica Multimediale, Carra per Informatica Generale).
6) Dopo la scadenza del 15/06 non sarà più possibile effettuare l'upload dell'elaborato. Chi non avrà consegnato perderà definitivamente il diritto di fare l'esame nella modalità orale.

Testo dell'elaborato

Scrivere un'applicazione C, sfruttando ove possibile le system call di Linux, per calcolare la frequenza di una parola con il paradigma MapReduce.

Si consideri un documento (formato testo) di grandi dimensioni. Data una parola chiave X, si vuole calcolare la frequenza media di tale parola all'interno del documento. In pratica si devono contare le occorrenze della parola X, n_X, e il numero totale di parole del documento, N_i, trovando la frequenza f_i definita come n_X / N_i.

Per risolvere tale problema, si utilizza il paradigma MapReduce.

Il programma principale inizia invocando il processo di Map. Durante tale fase, vengono instanziati P processi in parallelo che analizzano separatamente una porzione di documento. In pratica, il documento viene suddiviso in P porzioni non sovrapposte e dato in ingresso ai P processi di Map. Ciascun processo conta il numero di occorrenze della parola X e il numero di parole della porzione di documento assegnata. Come prima approssimazione, si possono ignorare le parole troncate durante il processo di divisione del file. Quando il processo ha finito di analizzare la porzione di documento assegnata, aggiorna due contatori posti in un'area di memoria condivisa: il primo contatore è il numero di occorrenze della parola X, il secondo contatore è il numero totale di parole. Si noti che i diversi processi aggiornano gli stessi due contatori, per cui è necessario prevedere un meccanismo di mutua esclusione regolato da un semaforo.

Quando tutti i P processi hanno finito, viene invocato il processo di Reduce. Durante tale fase, viene semplicemente calcolata la frequenza della parola X come rapporto tra il numero delle occorrenze di X e il numero delle parole. Si noti che il processo Reduce può essere invocato solo quanto tutti i

processi Map hanno terminato di aggiornare i due contatori, per cui bisogna sincronizzare il processo Reduce con la fine di tutti i processi Map attraverso un semaforo.

Una volta terminato di analizzare il documento, il programma dovrà creare un processo figlio Stampa, al quale dovrà spedire tramite *fifo* tutte le informazioni da stampare in un file. In particolare, Stampa dovrà stampare in un file il nome del documento analizzato, la parola X cercata, il numero delle occorrenze dalla parola X (ovvero n_X), il numero totale delle parole nel documento (ovvero N_i) e la frequenza f_i.

Nel risolvere il problema si tenga presente che:

- il documento da analizzare e la parola X da ricercare deve essere fornito nella linea di comando, così come il numero di processi P in parallelo;
- in ogni istante l'utente può richiedere il risultato parziale della frequenza dell'occorrenza della parola X (lo studente è libero di decidere la modalità con cui l'utente deve far terminare il server, ad esempio: CTRL+C, segnale esplicito tramite comando/system call *kill*, …).

Lo studente è libero di scegliere autonomamente come implementare caratteristiche del programma non espressamente specificate nelle regole precedenti ma necessarie al suo buon funzionamento.